Aritmética computacional

5. Representação de ponto flutuante

Ponto flutuante

- Representação de números não inteiros
  - Incluindo muito pequenos e muito grandes
- Parte da notação científica
  - -2.34×10^{56} ← normalizado
  - $+0.002 \times 10^{-4}$ ← não normalizado
  - $+987.02 \times 10^{9}$ ← não normalizado
- Em binário:
  - $\pm 1.xxxxxxx_2 \times 2^{yyyy}$
- Tipos `float` e `double` em C

Padrão de ponto flutuante

- Definido pela norma IEEE 754-1985
- Desenvolvido para padronizar as representações
  - Problemas de portabilidade para código científico
- Atualmente, é adotado mundialmente
- Duas representações
  - Precisão simples (32-bit)
  - Precisão dupla (64-bit)

Formato de ponto flutuante

	simples: 8 bits duplo: 11 bits	simples: 23 bits duplo: 52 bits
S	Exponent	Fraction

$$x = (-1)^S \times (1 + \text{Fraction}) \times 2^{(\text{Exponent} - \text{Bias})}$$

- S: bit de sinal (0 ⇒ não negativo, 1 ⇒ negativo)
- Significando normalizado: 1.0 ≤ |significando| < 2.0
  - Fração: o que aparece à direita do ponto
  - Significando: o número completo, incluindo o 1.
  - Sempre possui o bit 1 à esquerda do ponto binário, portanto este não precisa ser representado explicitamente (bit escondido)
- Expoente: representação por excesso: expoente verdadeiro + Bias
  - Garante que o expoente é sempre sem sinal
  - Precisão simples: Bias = 127; precisão dupla: Bias = 1023

Capacidade da precisão simples

- Expoentes 00000000 e 11111111 são reservados para uma representação especial
- Menor valor representável
  - Exponent: 00000001
    $\Rightarrow$ expoente verdadeiro = 1 – 127 = –126
  - Fração: 000…00 $\Rightarrow$ significando = 1.0
  - $\pm 1.0 \times 2^{-126} \approx \pm 1.2 \times 10^{-38}$
- Maior valor
  - expoente: 11111110
    $\Rightarrow$ expoente verdadeiro = 254 – 127 = +127
  - Fração: 111…11 $\Rightarrow$ significando ≈ 2.0
  - $\pm 2.0 \times 2^{+127} \approx \pm 3.4 \times 10^{+38}$

Capacidade da precisão dupla

- Expoentes 0000…00 e 1111…11 são reservados para uma representação especial
- Menor valor
  - Expoente: 00000000001
    $\Rightarrow$ expoente verdadeiro = 1 – 1023 = –1022
  - Fração: 000…00 $\Rightarrow$ significando = 1.0
  - $\pm 1.0 \times 2^{-1022} \approx \pm 2.2 \times 10^{-308}$
- Maior valor
  - Expoente: 11111111110
    $\Rightarrow$ expoente verdadeiro = 2046 – 1023 = +1023
  - Fração: 111…11 $\Rightarrow$ significando $\approx$ 2.0
  - $\pm 2.0 \times 2^{+1023} \approx \pm 1.8 \times 10^{+308}$

Precisão de um ponto flutuante

- Precisão relativa
  - Todos os bits da fração são significantes
  - Precisão simples: aprox. 2^{-23}
    - Equivalente a $23 \times \log_{10}2 \approx 23 \times 0.3 \approx 6$ casas decimais de precisão
  - Precisão dupla: aprox. 2^{-52}
    - Equivalente a $52 \times \log_{10}2 \approx 52 \times 0.3 \approx 16$ casas decimais de precisão

Exemplo de ponto flutuante

- Representar –0.75
  - $-0.75 = (-1)^1 \times 1.1_2 \times 2^{-1}$
  - S = 1
  - Fração = $1000\ldots00_2$
  - Expoente = –1 + Bias
    - Simples: –1 + 127 = 126 = 01111110_2
    - Dupla: –1 + 1023 = 1022 = 01111111110_2
- Simples: 1011111101000…00
- Dupla: 1011111111101000…00

Exemplo de ponto flutuante

- Que número é representado pelo ponto flutuante de precisão simples:

  11000000101000…00

  - S = 1
  - Fração = $01000\ldots00_2$
  - Expoente = $10000001_2 = 129$
- $x = (-1)^1 \times (1 + 01_2) \times 2^{(129 - 127)}$

  $= (-1) \times 1.25 \times 2^2$

  $= -5.0$